

Динамические тематические модели

Никита Дойков
МГУ

30 сентября 2014

Как учитывать время?

Время — дополнительная информация:

1. Улучшает качество модели;
2. Позволяет визуализировать результат.

Задачи анализа текстов:

новостные ленты, социальные сети, блоги, Твиттер;

Коллекция пресс-релизов министерств иностранных дел:

- ▶ $|D| = 2 \cdot 10^4$;
- ▶ Временной промежуток: 10 лет;
- ▶ Каждому документу соответствует временная метка:

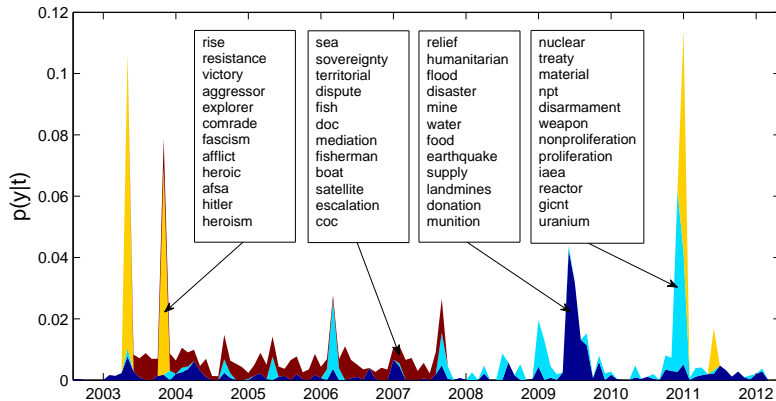
$$d \mapsto y_d \in Y,$$

Y — множество дней.

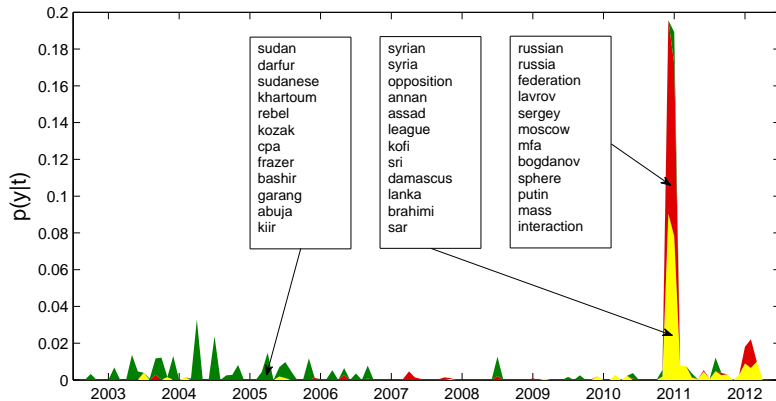
Число тем: $|T| = 100$.

- ▶ Европа — 11 тем;
- ▶ Америка — 14 тем;
- ▶ Африка — 10 тем;
- ▶ Западная Азия — 13 тем;
- ▶ Восточная Азия и Океания — 13 тем;
- ▶ Общие темы: Экономика, Наука, Свобода, Вооружение, ...

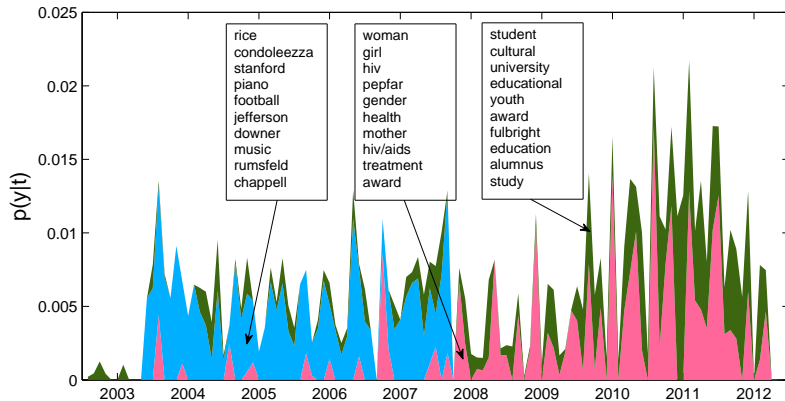
Результат



Результат



Результат



Предположения:

1. Распределение временных меток по документам — вырожденное:

$$p(y|d) = [y = y_d].$$

2. Гипотеза условной независимости:

$$p(y|t, d) = p(y|d).$$

Выводим распределения тем во времени:

$$p(y|t) = \frac{1}{p(t)} \sum_{d \in D_y} \theta_{td} p(d),$$

$$p(t|y) = \frac{1}{p(y)} \sum_{d \in D_y} \theta_{td} p(d).$$

Добавляем регуляризаторы времени

- **Регуляризатор 1:** в каждый момент времени число тематик невелико.

Разреживание $p(t|y)$ для всех $y \in Y$:

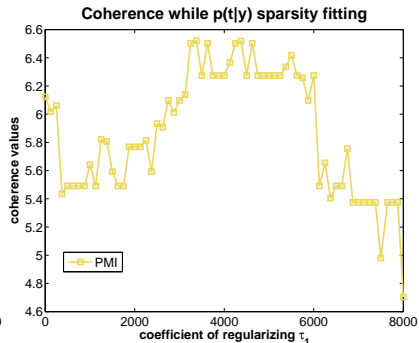
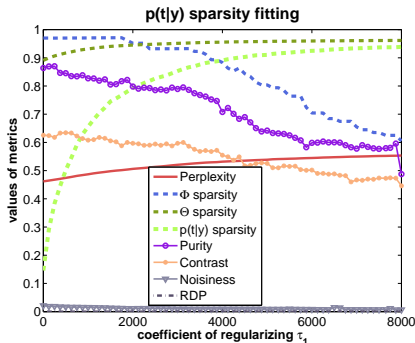
$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \log p(t|y) \longrightarrow \max.$$

- **Регуляризатор 2:** со временем темы должны меняться плавно, с редкими скачками.

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} |p(y|t) - p(y-1|t)| \longrightarrow \max.$$

Подбор коэффициентов регуляризации

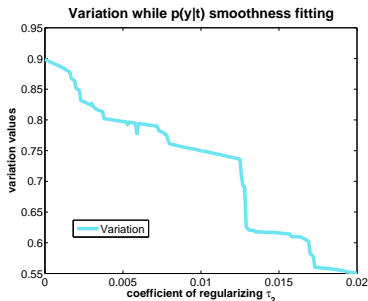
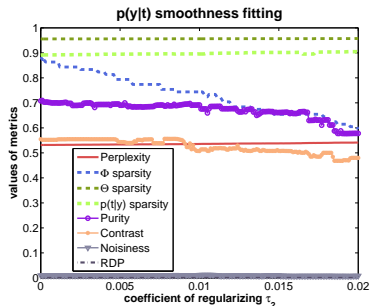
Разреживание $p(t|y)$:



- ▶ Регуляризатор разреживания $p(t|y)$ работает.
- ▶ Интерпретируемость увеличивается.

Подбор коэффициентов регуляризации

Сглаживание тем во времени:



► Колебание тем уменьшается.

$$\text{Variation}(t) = \sum_{y \in Y} \left| \sqrt{p(y|t)} - \sqrt{p(y-1|t)} \right|.$$

- ▶ Наши тематические модели работают!
- ▶ Время — дополнительная информация:
 1. Улучшает качество модели;
 2. Позволяет визуализировать результат.
- ▶ Интерпретируемые ограничения легко вносить в качестве регуляризаторов.